AFRL-OSR-VA-TR-2013-0576

(YIP-10) GEOMETRIC CLUSTERING AND ITS APPLICATIONS

PIYUSH KUMAR

FLORIDA STATE UNIVERSITY, THE

10/31/2013
Final Report

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)/RSL
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
10/29/2013	Final Technical Report	15-05-2010 - 14-05-2013

4. TITLE AND SUBTITLE

Geometric Clustering and its Applications

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-10-1-0230

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Piyush Kumar

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Florida State University
874 Traditions Way, 3rd Floor
Tallahassee, FL 32306-4166

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

USAF, AFRL, AFOSR
875 N. Randolph Street, Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Unclassified Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The AFOSR YIP Grant led to many activities and findings at FSU including the graduation of two PhD students who are now well placed. This report details the activities and findings of our project. The highlights of the achievements include: 1) Design of a new Support Vector Machine algorithm that is linearly convergent and yields an optimal number of support vectors. 2) A fast nearest neighbor algorithm and its optimized implementation in low dimensions. 3) A real time method to compute centers on maps given multiple addresses, when the distance is measured using shortest paths. 4) A 2-clustering algorithm for segments which is a pre-cursor to clustering curves that represent flight paths in air space. 5) A fast Euclidean MST algorithm and its implementation that computes clusterings with max spacing.

15. SUBJECT TERMS

Clustering, Optimization, Computational Geometry, Core-Sets

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	Unclassified Unlimited	6	Piyush Kumar
					19b. TELEPHONE NUMBER *(include area code)* 850-645-2355

Research Findings

1 Research Findings: Geometric Clustering and its applications

In a series of papers (jointly with collaborators) [14, 13, 12, 2, 10, 11, 8, 9, 18], I was one of the pioneers on solving geometric clustering problems using the approximation algorithms. In [10, 11], we showed the existence of $O(1/\epsilon)$ size core sets for the minimum enclosing ball (MEB) problem. For the minimum volume ellipsoid (MVE) problem, we have shown the existence of $O(d^2/\epsilon)$ size core-sets for $(1+\epsilon)$-approximation of the volume and $O(\frac{d \log d}{\epsilon})$ size core-sets for $(1+\epsilon)$ approximation of the radii [12]. Our algorithms for both of these problems are among the best available, both in terms of theoretical guarantees and practical results. Below I summarize recent findings that have been enabled by the **AFOSR YIP Grant** on various problems.

2 Support Vector Machines

We have recently improved the best core set algorithm [6] known for SVM problems [14]. We observed that one can remove the assumption of using an exact SVM Solver for the design of a core-set based algorithm to solve the SVM problem. Our algorithm outputs optimal number of support vectors (within a constant factor), and exhibits linear convergence. We also show implementation results, and compare our implementation with other major first order algorithms available. This algorithm is big data friendly, cache oblivious, easy to parallelize, apart from being very simple to implement.

3 New Approximate Nearest Neighbor Search Algorithm

We have implemented a fast, dynamic, approximate, nearest-neighbor search algorithm that works well in fixed dimensions ($d \leq 5$), based on sorting points (with integral coordinates) in morton (or z-) ordering [5]. Our code scal es well on multi-core/cpu shared memory systems. Our implementation is competitive with the best approximate nearest neighbor searching codes available on the web [17], especially for creating approximate k-nearest neighbor graphs of a point cloud. This is joint work with my graduate student Michael Connor [4].

An application of approximate nearest neighbor search is in the computation of Group Enclosing Queries [15]. Given a set of points P and a query set Q, a *group enclosing query* (GEQ) fetches the point $p^* \in P$ such that the maximum distance of p^* to all points in Q is minimized. For instance, given a large spatial database of points of interest, such as restaurants or resorts, a group of people trying to figure out a place to meet such that the longest distance traveled by anyone in the group is minimized, is an example of GEQ. This paper presents the challenges associated with such a query and proposes efficient, R-tree based algorithms for GEQ. If an exact answer is not critical, we present a simple and practical $\sqrt{2}$-approximation algorithm and extend it to retrieve $(1+\epsilon)$-approximate solutions for GEQ. Furthermore, our study on GEQ reveals its close relationship with the bichromatic *reverse furthest neighbors* problem (RFN), for which only limited theoretical treatment exists. As a by-product, we present the first R-tree based algorithm for RFN. Our algorithms do not assume that either P or Q fits in main memory. Experiments on both synthetic and real data sets confirm the superior efficiency and scalability of proposed algorithms over the naive, brute-force search based approach. We also made progress on the K-Nearest neighbor Joins in large relational databases [19].

3.1 Surface Reconstruction

In this part of the project, we use the implementation for finding fast approximate dynamic nearest neighbors to implement a fast, out of core, streaming, parallel, surface reconstruction algorithm. We have been experimenting with non-linear dimensionality reduction methods for use in surface reconstruction applications. We also show how our algorithm scales on multi-processor/core systems. This is joint work with my graduate student, James McClain.

On a related subject, we made progress on accurate localization of RFID tags in three dimensions [7].

4 Clustering on Road Networks

We studied the 1-center problem on road networks, an important problem in GIS. Using Euclidean embeddings, and reduction to fast nearest neighbor search, we devise an approximation algorithm for this problem. Our initial experiments on real world data sets indicate fast computation of constant factor approximate solutions for query sets much larger than previously computable using exact techniques [3, 16].

5 Bichromatic 2-Center of Pairs of Points

This study is motivated by a facility location problem in transportation system design, in which we are given origin/destination pairs of points for desired travel, and our goal is to locate an optimal road/flight segment in order to minimize the travel to/from the endpoints of the segment. We considered various variants of this problem, under different metrics and came up with efficient algorithms, both approximation and exact during this research [1]. Most of the linear or near-linear time algorithms found during this research were inspired by the core-set approach.

References

[1] Esther M. Arkin, Jos Miguel Daz-Bez, Ferran Hurtado, Piyush Kumar, Joseph S. B. Mitchell, Beln Palop, Pablo Prez-Lantero, Maria Saumell, and Rodrigo I. Silveira. Bichromatic 2-center of pairs of points. In David Fernndez-Baca, editor, *LATIN*, volume 7256 of *Lecture Notes in Computer Science*, pages 25–36. Springer, 2012.

[2] Y. Bulatov, S. Jambawalikar, P. Kumar, and S. Sethia. Hand recognition using geometric classifiers. In *Proceedings of International Conference on Biometric Authentication*, volume 3072 of *Lecture Notes Comput. Sci.*, pages 753–759. Springer-Verlag, 2004.

[3] Samidh Chatterjee, Bradley Neff, and Piyush Kumar. Instant approximate 1-center on road networks via embeddings. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 369–372, New York, NY, USA, 2011. ACM.

[4] Michael Connor and Piyush Kumar. Parallel construction of k-nearest neighbor graphs for point clouds. In *Proceedings of Volume and Point-Based Graphics.*, pages 25–32. IEEE VGTC, August 2009.

[5] Michael Connor and Piyush Kumar. Practical nearest neighbor search in the plane. In Paola Festa, editor, *Experimental Algorithms*, volume 6049 of *Lecture Notes in Computer Science*, pages 501–512. Springer Berlin / Heidelberg, 2010.

[6] S. Har-Peled, D. Roth, and D. Zimak. Maximum margin coresets for active and noise tolerant learning. Technical Report No. UIUCDCS-R-2006-2784, UIUC Computer Science Department, Oct 2006.

[7] C. Hekimian-Williams, B. Grant, Xiuwen Liu, Zhenghao Zhang, and P. Kumar. Accurate localization of rfid tags using phase difference. In *RFID, 2010 IEEE International Conference on*, pages 89–96, 2010.

[8] Sachin Jambawalikar and Piyush Kumar. A note on approximate minimum volume enclosing ellipsoid of ellipsoids. *International Conference on Computational Sciences and Its Applications*, 0:478–487, 2008.

[9] P. Kumar and P. Kumar. Almost optimal solutions to k-clustering problems. Unpublished Manuscript, Submitted to IJCGA.

[10] P. Kumar, J. S. B. Mitchell, and A. Yıldırım. Approximate minimum enclosing balls in high dimensions using core-sets. *The ACM Journal of Experimental Algorithmics*, 8, 2003.

[11] P. Kumar, J. S. B. Mitchell, and A. Yıldırım. Computing core-sets and approximate smallest enclosing hyperspheres in high dimensions. In *Algorithm Engineering and Experimentation (Proc. ALENEX '03)*, Lecture Notes Comput. Sci., pages 45–55. Springer-Verlag, 2003.

[12] P. Kumar and A. Yıldırım. Minimum volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and Applications*, 126(1):1–21, 2005. `http://www.ams.sunysb.edu/~piyush/papers/mve.pdf`.

[13] P. Kumar and A. Yıldırım. Computing minimum volume enclosing axis-aligned ellipsoids. *Journal of Optimization Theory and Applications*, 136(2):211–228, 2008.

[14] P. Kumar and E. A. Yıldırım. A linearly convergent algorithm for support vector classification with a core-set result. *INFORMS Journal on Computing*, 2010. To Appear.

[15] Feifei Li, Bin Yao, and Piyush Kumar. Group enclosing queries. *IEEE Trans. on Knowl. and Data Eng.*, 23(10):1526–1540, October 2011.

[16] James McClain and Piyush Kumar. Fast k-clustering queries on embeddings of road networks. In *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*, COM.Geo '12, pages 11:1–11:9, New York, NY, USA, 2012. ACM.

[17] D. Mount. ANN: Library for Approximate Nearest Neighbor Searching, 1998. `http://www.cs.umd.edu/~mount/ANN/`.

[18] P.Kumar and A. Yildirim. An algorithm and a core set result for the weighted euclidean one-center problem. *Informs Journal on Computing*, 2009. To Appear.

[19] Bin Yao, Feifei Li, and P. Kumar. K nearest neighbor queries and knn-joins in large relational databases (almost) for free. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 4–15, 2010.

A. PROJECT ACTIVITIES

1 Project Summary

The primary objective of this YIP proposal is the design, analysis and implementation of efficient algorithms for geometric clustering problems and their applications. Core-Sets are ideal for many such designs. Using this paradigm, one computes a small but 'most relevant' subset of the input, solves the optimization problem on this small subset thereby calculating an approximate solution to the original problem with proven accuracy and efficiency. The main focus here will be on the design, analysis, and implementation of efficient algorithms for problems arising in computational geometry and discrete optimization.

The educational component of this project includes reworking the algorithms and computational geometry courses, disseminating teaching material over the web, directing undergraduate and graduate student and software development integration into education.

2 Activites

Research activities have been undertaken for each component of this project. We report the advances we have made in the following subsections (in terms of observations, simulations, experiments and presentations).

2.1 Geometric Clustering on Maps

One of the applications we worked on is the computation of centers for Geographic Information System applications. We built algorithms and software that would be able to compute meeting points for multiple people/addresses accurately and in real time. Work on this project has begun and a very early prototype can be found at: http://maps.compgeom.com We also published this work in ACM SIGSPATIAL 2011. This work was later extended to allow multiple center computations in real time. This work also won the NSF ICorps Award in 2012.

2.2 Approximate nearest neighbor search

In this part of the project, we have implemented and experimented with a fast dynamic approximate nearest neighbor search algorithm that works well in fixed dimensions ($n \leq 5$). Our algorithm scales well on multi-core / multi-cpu shared memory systems. Preliminary experiments showed that our method is competitive with the best approximate nearest neighbor searching codes available on the web (ANN by David Mount). A preliminary software release related to this project can be found at: http://compgeom.com/~stann

Applications to Spatial Databases: Finding the k nearest neighbors (kNN) of a query point, or a set of query points (kNN-Join) are fundamental problems in many database application in the Air Force. Most of the previous efforts to solve these problems focused on spatial databases or stand-alone systems, where changes to the database engine are required, which greatly limits their application on large data sets that are stored in a relational database management system. Furthermore, these methods cannot automatically optimize kNN queries or kNN-Joins when additional query conditions are specified. In this work, we study both the kNN query and the kNN-Join in a relational database, possibly augmented with additional query conditions. We search for relational algorithms that require no changes to the database engine. The straightforward solution uses the user-defined-function (UDF) that a query optimizer cannot optimize. We design algorithms that could be implemented by SQL operators without changes to the database engine, hence enabling the query optimizer to understand and generate the best query plan. Extensive experiments on large, real and synthetic, data sets confirm the superior efficiency and practicality of our approach, compared to the state of the art.

Applications to Geometric Minimum Spanning Tree: We have recently devised an algorithm that can compute GMSTs on multi-core machines. We call it GeoFilterKruskal, an algorithm that computes the minimum spanning tree of a set of points P, using well separated pair decomposition in combination with a simple modification of Kruskal's algorithm. When P is sampled from uniform random distribution, we show that our algorithm runs in $O(n \log^2 n)$ time with high probability. Experiments show that our algorithm works better in practice for most data distributions compared to the current state of the art. Our algorithm is easy to parallelize and to our knowledge, is currently the best practical algorithm on multi-core machines for dimensions greater than two. The GMST algorithm was released as an open source code to the public as part of the STANN library.

Fast Exact nearest neighbors in 2D: In 2010, we showed that using some very simple practical assumptions, one can design an algorithm that finds the nearest neighbor of a given query point in $O(\log n)$ time in theory and faster than the state of the art in practice. The algorithm and proof are both simple and the experimental results clearly show that we can beat the state of the art on most distributions in two dimensions.

2.3 Surface Reconstruction

In this part of the project, we use the implementation for finding fast approximate dynamic nearest neighbors to implement a fast, out of core, streaming, parallel, surface reconstruction algorithm. We have also been experimenting with non-linear dimensionality reduction methods for use in surface reconstruction applications. We have begun to get results for non-smooth surfaces in this area and hope to report our results soon to leading conferences.

A related endeavor we embarked on is accurate localization of RFID tags in three dimensions. We have also made progress on this problem.

2.4 Core-Sets: Support Vector Machines

We have recently improved the best core set algorithm known for SVM problems. We observed that one can remove the assumption of using an exact SVM Solver for the design of a core-set based algorithm to solve the SVM problem. We are currently in the process of implementing our algorithm.

2.5 Bichromatic 2-Center clustering for pairs of points

We study a class of geometric optimization problems closely related to the 2-center problem: Given a set S of n pairs of points, assign to each point a color (red or blue) so that each pairs points are assigned different colors and a function of the radii of the minimum enclosing balls of the red points and the blue points, respectively, is optimized. In particular, we consider the problems of minimizing the maximum and minimizing the sum of the two radii. For each case, minmax and minsum, we consider distances measured in the L_2 and in the L_∞ metrics. Our problems are motivated by a facility location problem in transportation system design, in which we are given origin/destination pairs of points for desired travel, and our goal is to locate an optimal road/flight segment in order to minimize the travel to/from the endpoints of the segment.

2.6 Education

The PI is continuously integrating research findings from this project into his graduate courses. The PI has completely reworked and developed a new curriculum for his graduate course on computational geometry. The homework assignments in this course involve computational projects that require each student to formulate and code geometric optimization problems. This experience has been extremely valuable as many graduate students continued to use core-set techniques in their own research.

Since the PIs interest are both in theory and implementation, he recently moved his group to the Python programming language and has designed a new course on the subject so that not only his group, but the students at FSU can benefit from such a course. The Python course that he designed covers not only the basics of Python but more advanced material like metaprogramming and decorators in detail. In the near future, the PI believes that this will help the students get things done much faster than the students coding in C++.

The AFOSR Funding helped two PhD Students to graduate and get well placed. Michael Connor now works for NSA and Samidh Chatterjee works in a startup named AirSage that processes large data mobile paths. Two more PhD students have availed this funding and are close to graduation.

2.7 Presentations

The PI and his students made the following presentations during the course of the project, till date (starting March 2010):

- *K Nearest Neighbor Queries and KNN-Joins in Large Relational Databases (Almost) for Free.* At 26th IEEE International Conference on Data Engineering. Long Beach, California, USA , March 2010.
- *Nearest neighbors in the Plane.* At Brooklyn Polytecnic, NY. April 2010.
- *Accurate Localization of RFID Tags Using Phase Difference.* At IEEE RFID, Orlando, FL. April 2010.
- *Nearest neighbors in the Plane.* At Symposium on Experimental Algorithmics, Napoli, Italy. May 2010.
- *Geometric Minimum Spanning Trees with GeoFilterKruskal.* At Symposium on Experimental Algorithmics, Napoli, Italy. May 2010.
- *Core-sets and its applications.* REEF, Destin, FL. Aug 2010.
- *Minimum Error Rate Training by Sampling the Translation Lattice.* At 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010). Boston, MA. October 2010.
- *Instant Approximate 1-Center on Road Networks Via Embeddings.* At ACM SIGSPATIAL GIS, 2011.
- *Instant Approximate 1-Center on Road Netwroks Via Embeddings.* At TAMU, College Station, TX, October 2011.
- *Clustering Large Data sets.* At BP Research, Houston, TX, Aug 2011.
- *Fast k-clustering Queries on Road Networks.* At Com.Geo 2012, Reston, VA, 2012.
- *Bichromatic 2-Center of Pairs of Points.* At LATIN 2012, Arequipa, Peru, 2012.
- *Fast Multiple center map queries* At NSF ICorps, Ann Arbor, 2012.